Semantic segmentation of urban scenes by learning local class interactions

Semantic segmentation of urban scenes by learning local class interactions

Michele Volpi Vittorio Ferrari
University of Edinburgh
{v1mvolpi,vferrari}@staffmail.ed.ac.uk

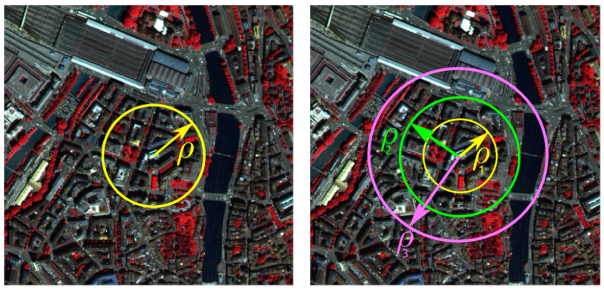
Abstract

Traditionally, land-cover mapping from remote sensing images is performed by classifying each atomic region in the image in isolation and by enforcing simple smoothing priors via random fields models as two independent steps. In this paper, we propose to model the segmentation problem by a discriminatively trained Conditional Random Field (CRF). To this end, we employ Structured Support Vector Machines (SSVM) to learn the weights of an informative set of appearance descriptors jointly with local class interactions. We propose a principled strategy to learn pairwise potentials encoding local class preferences from sparsely annotated ground truth. We show that this approach outperform standard baselines and more expressive CRF models, improving by 4-6 points the average class accuracy on a challenging dataset involving urban high resolution satellite imagery.

1. Introduction

Segmenting very high resolution (VHR) satellite or aerial images into semantic regions is one of the most active research areas in the remote sensing community [12, 18, 25, 27, 30]. Providing high quality land-cover maps is of crucial importance to many fields of environmental science, ranging from deforestation analysis to urban modeling [1, 23]. VHR imagery allows to obtain such maps with unprecedented spatial detail. In this paper, we consider a semantic segmentation problem involving VHR satellite scenes of urban areas with sparse annotations.

For this data, the high spatial variability of the spectral signal does not directly correspond to changes in the label field. To cope with these adverse effects, one has to consider the urban class appearance jointly with their spatial arrangement (local context) via structured output models. While learning the relationship between the observed data and the label can be carried out by combining informative descriptors and powerful classifiers, it is less clear how to efficiently learn local class interactions from sparsely annotated ground truth. The central contribution of this work is a principled scheme for structured learning of local spatial interactions in such conditions.



(a) (b)

Figure 1. **Learning local class interactions with rings:** (a) single ring potential (SRP) or by (b) multiple rings potentials (MRP).

Markov Random Fields (MRF) [2] and Conditional Random Fields (CRF) [16] are structured output models that allow accounting for interactions of random variables. Such methods elegantly model cues from the image data along with prior beliefs about output dependencies and their spatial organization. These approaches have been successfully exploited in both the remote sensing [13, 12, 18, 30] and the computer vision communities [10, 11, 15, 17, 19]. However, little effort has been put in *learning* such models specifically tailored for remote sensing data. A possible reason behind the scarce success of CRF learning in remote sensing applications may be due to the sparsity of ground truth annotations. These images are characterized by high ambiguity in appearance and therefore manual annotation of full scenes is extremely costly and time consuming, often involving ground surveys.

In this paper, we propose a new strategy to learn CRF parameters (Sec. 2) based on ring-based class-interaction potentials (Sec. 3). Since all georeferenced satellite or aerial images share the same underlying geographical space (i.e. *geographic coordinates*) we can exploit these data-specific regularities to learn discriminative CRF parameters by structured support vector machines (SSVM, Sec. 4). This strategy overcomes the problems introduced by sparse annotations during structured learning of CRF parameters. Experiments on satellite VHR urban semantic segmentation (Sec. 5) show that learning the weights jointly combining appearance descriptors and spatial class preferences improves significantly standard approaches.

1.1. Related work and motivation

Random fields models in remote sensing. Standard MRF and CRF have been successfully applied to remote sensing image labeling tasks. MRF have been employed for segmentation of both synthetic aperture radar and optical images [28, 18, 22]. These models usually enforce smoothness priors, reducing the effects of the speckle noise (in radar imaging) or high resolution (optical data). For optical images, Schindler [22] observed that spatially smooth labeling may significantly increase the segmentation accuracy of VHR images, when comparing to classification of atomic regions of the image in isolation.

More flexible CRF models have also been employed. Kluckner *et al.* [13] utilized covariance descriptors to classify the appearance of pixels and then combined them into a CRF enforcing smoothness over adjacent regions with similar elevation ranges. They also include co-occurrence by counting class associations on the training set. Hoberg *et al.* [12] extended contrast-sensitive CRF to include temporal dependencies between images acquired at different time instants over the same geographical area. However, in these works, no structured learning of CRF parameters is performed. Tuia *et al.* [27] presented a SSVM approach to pixel classification. However, rather than accounting for the spatial dependence between random variables, they consider a pre-defined spectral class taxonomy through a hierarchical loss. Zhong *et al.* [31] employed maximum likelihood training to find parameters of a set of contrast sensitive CRF, one for each type of descriptor. Then, they combined the CRF models to extract urban areas from multisensor images acquired. However, they focused on fusion of simple CRF models involving binary problems. More recent trends to CRF-based segmentation aim at including higher-order terms to enforce smoothness among group of atomic regions for specific urban classes. Wegner *et al.* [30] adopted a robust P^n-Potts potential to extract road networks, with cliques composed by road candidates.

In general, structured output models employed so far in satellite or aerial image semantic segmentation rely on generic predefined smoothing terms (e.g. contrast-sensitive Potts) or on very specific adaptations as dictated by the application at hand (e.g. class-specific high-order potentials for road extraction). Although the apparent benefits of such models, no further developments in structured learning of CRF models have been proposed for this specific kind of data.

Random fields models in computer vision. MRF and CRF are ubiquitous in computer vision applications. Their success is acknowledged in many problems, in particular on semantic segmentation tasks [20]. An interesting family of CRF models jointly the appearance of classes with pairwise terms enforcing both contrast-sensitive smoothness and class co-occurrences [21]. For instance, Ladický *et al.* [15] modeled co-occurrence between classes in training images and exploited these regularities as an additional high-order image-dependent term in a contrast-sensitive CRF. Results show that including these additional cues increase the accuracy by avoiding unlikely label combinations. Other approaches exploiting co-occurrence [21, 9, 24] enforced class preferences through pairwise random field models. Lucchi *et al.* [17] analyzed the effects of local and global (higher-order) constraints for semantic segmentation using CRF learned by SSVM. Their models implemented various spatial constraints and also included learned label interactions constrained to particular cardinal directions. A similar line of reasoning supplements segmentation models by a data-driven prior encoding the relative location and frequency between classes [10, 11].

Note that these models employ the available ground truth in different ways. On the one hand, a full pixel-level annotation is not required to learn coarse co-occurrence potentials (e.g. at the image level as in [15] or "above to"-"below to" and "right-to"-"left-to" as in [17]). On the other hand, one needs full pixel level annotations to properly learn class preferences between nodes connected in a standard adjacency graph.

Our strategy to learn context. In this work we assume that a sparse but homogeneous ground truth is available. In our setting, ring-based potentials allow learning of class preferences in a locally isotropic manner, an assumption satisfied in aerial urban scenes. Class co-occurrences at the scene level does not bring any additional information, since different urban classes are likely to occur uniformly. On the opposite situation, learning class correspondence directly on adjacent nodes may be suboptimal, since only very particular class associations may be represented. For instance, urban classes such as "Trees" and "Gardens" may lie between classes such as "Buildings" and "Roads", which intuitively should co-occur more frequently as direct neighbors in a urban context. Moreover, labeled nodes which are direct neighbors are often biased to same class assignments, due to the nature of the annotations. For these reasons, we develop a more robust but locally discriminative solution relying on ring potentials, properly exploiting the *geographical* space. By this approach, both "Building"-"Roads" and "Building"-"Gardens" will be frequently observed together. At the same time, this modeling strategy allows to efficiently cope with sparse annotations. In this paper we consider two learning strategies. The first, simpler, considers local circles with a pre-defined radius as depicted in Fig. 1(a). The second extends this line of reasoning by accounting simultaneously for different concentric rings alleviating the manual definition of a single radius and showing a more flexible learning strategy (Fig. 1(b)).

2. CRF for semantic segmentation

Our pairwise CRF model is composed by unary and pairwise terms which jointly describe interactions between input and output variables. In our setting, we are given a set of N *training images* $X = \{\mathbf{x}^n\}_{n=1}^N \in \mathcal{X}$ with corresponding *ground truth annotations* $Y = \{\mathbf{y}^n\}_{n=1}^N \in \mathcal{Y}$. Unlike standard supervised classification, the labelings are not isolated discrete values but structured objects. As such, the labeling $\mathbf{y}^n$ is a configuration of labels y_i assigned to each superpixel (node) x_i on the image plane. These dependencies are described by an irregular graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes connected by undirected edges $\mathcal{E}$. Usually, superpixels sharing some boundary are linked by an edge.

The optimal labeling of the image is found by minimizing an energy function of the form:

$$E(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \sum_{i \in \mathcal{V}} \varphi_i(y_i; \mathbf{w}^{\varphi}) + \sum_{(ij) \in \mathcal{E}} \phi_{ij}(y_i, y_j; \mathbf{w}^{\phi}). \tag{1}$$

The terms $\varphi_i(y_i)$ and $\phi_{ij}(y_i, y_j)$ are respectively the unary and pairwise potentials, both depending on node labels y_i, y_j, on the parameter vectors $\mathbf{w}^{\varphi}$, $\mathbf{w}^{\phi}$ learned by SSVM and on some image evidence x_i, x_j (the latter dropped from both potentials in Eq. (1) for clarity purposes).

This inference problem can be solved with standard graph-cuts solvers when the energy is submodular [4] or by QPBO graph-cuts when it is not [14]. We will discuss model inference in more detail in Section 4.1.

3. Energy model

The weights of unary and pairwise potentials in the energy function in Eq. (1) are jointly learned by SSVM [26]. This strategy allows the parameters to better adapt to the training data by putting weight to the mapping functions that are important in separating the ground truth labeling $\mathbf{y}^n$ from wrong labelings $\mathbf{y}$ with low energy. This ultimately provides better segmentation accuracy and results in a more elegant strategy than, for instance, cross-validating a trade-off parameter between precomputed potentials.

In this Section, we present the building blocks of the energy function in Eq. (1) and our strategy to learn its parameters. In particular, we present a principled strategy relying on spatial rings to perform strucured learning.

3.1. Unary potentials

The unary potential measures the likelihood that a node takes a particular label, based on its appearance. It is commonly the probabilistic output of a discriminative appearance classifier. The potential is usually defined by taking the negative log-likelihood as $\varphi_i(y_i) = -\log\big(p(y_i|x_i)\big)$. However, note that in our setting the SSVM directly learns a weighted linear combination of input descriptors. By defining $\psi_i(y_i)$ as the descriptor set corresponding to node i for class y_i, SSVM learns $\mathbf{w}^{\varphi}$ so that $\varphi_i(y_i) = \langle \mathbf{w}^{\varphi}, \psi_i(y_i) \rangle$ is a linear classifier.

3.2. Pairwise potentials

Pairwise potentials are designed to encode our prior belief about relationships between random variables. The most commonly employed prior is the Potts smoothing, which encourages adjacent nodes to share the same label by $\phi_{ij}^p(y_i, y_j) = [\![y_i \neq y_j]\!]$, where $[\![y_i \neq y_j]\!]$ returns 1 when $y_i \neq y_j$ and 0 otherwise. This potential can be made contrast-sensitive by including a term adapting to the appearance of the connected nodes. The contrast sensitive Potts potential is defined as $\phi_{ij}^s(y_i, y_j) = g(x_i, x_j)[\![y_i \neq y_j]\!]$, where g is a function estimating the similarity of the superpixels i and j based on the appearance descriptors x_i and x_j. In this case, the pairwise term encourages a label switch if the two superpixels are different under g.

The form of a standard co-occurrence pairwise potentials is $\phi_{ij}^c(y_i, y_j) = h(y_i, y_j)[\![y_i \neq y_j]\!]$ [24]. The function h estimates a preference score between y_i and y_j. These potentials can be learned by counting label occurrencies in the training data and will encourage outputs with common class associations, while discouraging rare class co-occurrences. Co-occurrence potentials can be combined with the contrast-sensitive Potts potential to account for appearance differences. Dealing with remote sensing data, co-occurrence interactions have to be estimated locally for each node. As mentioned in the introduction, most of the classes are likely to co-occur uniformly at the image level, but they cannot be learned directly from first order neighborhoods because of sparseness and bias in the ground truth.

Single ring potentials (SRP). In this work, we introduce ring potentials to learn discriminatively the contrast-sensitive Potts potential weights. These weights directly indicate whether particular urban class associations should be favored or disfavored by the model. To this end, pairwise potentials are learned from regions defined by a circle centered on each superpixel, as illustrated in Fig. 1(a). We make the assumption that the ring has to be large enough to compensate for the spatially sparse annotations, but at the same time has to be small enough to depict local characteristics of the label field, thus not being too general. We let the SSVM learn the weights w_{y_i,y_j} of a pairwise potential formulated as $\phi_{ij}(y_i, y_j) = w_{y_i,y_j} g(x_i, x_j)[\![y_i \neq y_j]\!]$.

During learning, undirected edges $ij \in \mathcal{E}^{\rho}$ connect all the nodes j contained in a circle of radius ρ (in meters) centered on i (ρ-ball spatial graph). Therefore, the value of the weight w_{y_i,y_j} directly discriminates class interactions between y_i and y_j, locally. Also, note that the potential includes contrast sensitivity between pairs of nodes as defined by g. By selecting an appropriate radius ρ around node i, the SSVM can learn meaningful local interactions.

Multiple rings potential (MRP). The optimal size of the SRP is not known a priori and its selection may involve a costly cross-validation process. For this reason, we propose an additional strategy allowing a joint learning of pairwise potentials for *multiple concentric rings* to relax the SRP. An illustration is given in Fig. 1(b).

This strategy is a generalization of the SRP, which allows optimizing different spatial interaction potentials in multiple rings at a time. To this end, we introduce a set of concentric rings defined by a set of radii $\rho = \{\rho_e\}_{e=1}^{R}$. Each node i is connected to nodes j which are located at R different levels. Then, we learn $w_{y_i,y_j}^{\rho_e}$ for each ring. This way, not only the selection of the optimal single radius size is alleviated, but the model will be also able to learn optimal class preferences at multiple spatial buffers. The formulation of the pairwise potential is the same as for the SRP and only the connectivity graph varies according to $\mathcal{E}^{\rho_e}$. Specifically, $ij \in \mathcal{E}^{\rho_e}$ iff $\rho_{e-1} < d(x_i, x_j) \leq \rho_e$, where d returns the geographical (Euclidean) distance between the coordinates of superpixels centers on the image plane.

Practically, the SRP SSVM learns a $|\mathcal{Y}| \times |\mathcal{Y}|$ pairwise weight matrix, where $|\mathcal{Y}|$ indicates the number of classes. In the case of MRP, SSVM learns such matrix for each interaction level, therefore MRP counts $R|\mathcal{Y}|^2$ class interaction parameters. In both cases, potentials can be rewritten as the linear combination $\phi_{ij}(y_i, y_j) = \langle \mathbf{w}^{\rho_e}, \psi_{ij}(y_i, y_j) \rangle$, by letting $\psi_{ij}(y_i, y_j) = g(x_i, x_j)[\![y_i \neq y_j]\!], \forall\, ij \in \mathcal{E}^{\rho_e}$.

4. Max-margin structured learning

The goal of max-margin structured learning is to learn discriminatively the weights of the energy function in Eq. (1). To learn the parameters of CRF via SSVM, we must rewrite the energy so that it results in a linear combination of weights $\mathbf{w}$ and mapping functions Ψ as:

$$E(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle, \tag{2}$$

In our setting, the potentials employed are linear with respect to the parameter vector. This allows us to rewrite the energy as an inner product of weights and mapping functions. Specifically, the energy in Eq. (1) corresponds to a sum of potentials, which implies that they can be rewritten as $\varphi_i(y_i) = \langle \mathbf{w}^{\varphi}, \psi_i(y_i) \rangle$ and $\phi_{ij}(y_i, y_j) = \langle \mathbf{w}^{\phi}, \psi_{ij}(y_i, y_j) \rangle$. We now let $\Psi^{\varphi}(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \psi_i(y_i)$ and $\Psi^{\phi}(\mathbf{x}, \mathbf{y}) = \sum_{(i,j) \in \mathcal{E}^{\rho}} \psi_{ij}(y_i, y_j)$. We obtain the energy function expressed in Eq. (2) by concatenating mapping functions and corresponding weights $\Psi(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \Psi^{\varphi}(\mathbf{x}, \mathbf{y}) \\ \Psi^{\phi}(\mathbf{x}, \mathbf{y}) \end{bmatrix}$ and $\mathbf{w} = \begin{bmatrix} \mathbf{w}^{\varphi} \\ \mathbf{w}^{\phi} \end{bmatrix}$.

To learn the CRF parameters $\mathbf{w}$ we adopt the margin rescaling variant of the SSVM [26]. The intuition behind this approach is to find model weights that maximally separate the energy of any labeling $\mathbf{y}$ to the one of the ground truth $\mathbf{y}^n$ by the largest margin $\Delta(\mathbf{y}, \mathbf{y}^n)$. The maximization of the margin provides regularization allowing good generalization to previously unseen test images.

The estimation of $\mathbf{w}$ can then be formulated as a standard quadratic problem of the form [26]:

$$\begin{aligned} &\min_{\mathbf{w}, \boldsymbol{\xi} \geq \mathbf{0}} \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{N}\sum_{n=1}^{N} \xi_n \\ &\text{s.t. } \forall n, \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}^n : \\ &E(\mathbf{x}^n, \mathbf{y}; \mathbf{w}) - E(\mathbf{x}^n, \mathbf{y}^n; \mathbf{w}) \geq \Delta(\mathbf{y}, \mathbf{y}^n) - \xi_n, \end{aligned} \tag{3}$$

where ξ are slack variables allowing to solve non separable training data, C is a user defined penalization hyperparameter and $\Delta(\mathbf{y}, \mathbf{y}^n)$ is a loss function measuring the disagreement between $\mathbf{y}$ and $\mathbf{y}^n$. Note that the optimization problem in Eq. (3) is not directly tractable due to the exponential number of constraints, one for each possible configuration of $\mathbf{y}$. To this end, Tsochantaridis *et al.* [26] propose to solve the problem by cutting planes, which iteratively update the set of active constraints and solves the problem on this reduced working set. The most violated constraints are found by loss-augmented inference as:

$$\bar{\mathbf{y}} = \arg\min_{\mathbf{y} \in \mathcal{Y}} \; E(\mathbf{x}^n, \mathbf{y}; \mathbf{w}) - \Delta(\mathbf{y}, \mathbf{y}^n) \tag{4}$$

By employing a loss function which decomposes over nodes $i \in \mathcal{V}$, the above problem may be reduced to a standard MAP inference problem. This can be seen as augmenting the random field with an additional unary potential, representing the cost of nodes mislabeling. This way, the optimization focuses on weights separating ground truth from similar low energy labelings. In this work we employ the common Hamming loss, which is decomposable over the nodes. It penalizes wrongly labeled examples equally as $\Delta(\mathbf{y}, \mathbf{y}^n) = \frac{1}{|\mathcal{V}|}\sum_{i \in \mathcal{V}}[\![y_i \neq y_i^n]\!]$,

4.1. Inference

One has to solve an inference problem to find both the final labeling at test time in Eq. (2) and the most violated constraints in Eq. (4). In the case of binary labeling problems and submodular energy functions, this can be solved exactly and efficiently in polynomial time by using st-mincut/maxflow graph cuts algorithm [4]. However, we may encounter non-submodular energy functions when the CRF parameters are learned without imposing additional constraints, such as non-negativity of the pairwise weights. For this reason, we solve the graph-cuts problem using the Quadratic Pseudo-Boolean Optimization (QPBO) [3, 14]. This approach returns a node labeling $y_i = \{0, 1, \emptyset\}$, depending on the strength of the submodularity. Labeled nodes are guaranteed to be optimal. To solve multi-class inference problems, we further include an additional approximation by formulating the problem with α-expansions

moves [4]. This approach decomposes the multi-label problem in a series of binary energy minimization problems, by allowing a node to keep its current state or to switch to state α if the move decreases the energy. Since in our situation QPBO returns very few unlabeled nodes, we heuristically decided not to switch the label to α for unlabeled nodes and maintaining their current state.

Inference with learned rings potentials. For inference with ring potentials we consider a different graph connectivity than the one used for learning, to efficiently assimilate the learned spatial interactions into the standard pariwise CRF model. The new connectivity simply connects neighboring superpixels sharing some border. In the case of SRP, we assign to each edge the learned w_{y_i,y_j}. In the case of MRP, the value assigned to the edge depends on the distance between the center of mass of adjacent superpixels, as $w^{\rho_e}_{y_i,y_j}$ if $\rho_{e-1} < d(x_i,x_j) \leq \rho_e$. By employing this apparently discrepant strategy with respect to the learning step, we can reduce the size of the inference problem while exploiting efficiently learned class interactions.

5. Experiments

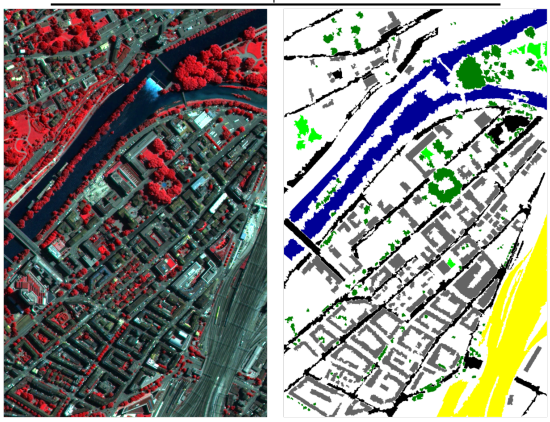
5.1. Setup

We tested the proposed system on a set of 20 multispectral VHR images acquired over the city of Zurich (Switzerland) by the QuickBird satellite in 2002. The average image size is 1000×1150 pixels (approximately 23M pixels in total) and they are composed of 4 channels spanning near-infrared to visible spectrum (NIR-R-G-B). The spatial resolution of the images after pan-sharpening is 0.61 meters / pixel. We manually annotated 8 different urban classes by labeling regions for which we were able to confidently identify the correct class. An example along with the urban class legend is shown in Fig. 2.

We transformed the images into a superpixel representation by the method of [8]. This allowed us to reduce the size of the problem and to provide a spatially coherent support for more advanced appearance descriptors. After transformation, we have 53k labeled superpixels out of a total of 113k. This approach is very appropriate for multi-spectral images, since it relies on distances computed on the whole spectrum and not on specific color spaces. Note that this method does not constraint directly the size of superpixels, and consequently two large adjacent superpixels may have a considerable relative distance between their centers. Our ring-based approach directly deals with these situations. We selected the parameters of the superpixel generator by empirically optimizing a trade-off between accuracy and number of superpixels, similarly to [19].

To solve the max-margin problem in Eq. (3) we employed the MATLAB interface to SVM$^{\text{struct}}$ provided by A. Vedaldi [29]. To solve the inference problems we used a MATLAB interface of the QBPO software by Kolmogorov *et al.* [14][1].

ID	Color	Label	ID	Color	Label
1		Roads	2		Buildings
3		Trees	4		Grass
5		Bare Soil	6		Water
7		Rails	8		Pools



(a) (b)

Figure 2. **Example from the Zurich dataset.** Original image (NIR-R-G) (a) and its ground truth (b). For the class legend, see the table above. Note that white background is not considered as a separate class.

Unary potentials. Feature functions $\psi_i(y_i)$ are composed by a set of informative descriptors computed for each superpixel. The appearance of each node is described by a combination of spectral information, normalized difference vegetation index, 8 sets of bag-of-visual-words descriptors from 4 different filters (300 words each with BoW extracted from the original superpixel size and a BoW from the same superpixel plus 15 pixels buffer to include context), location and superpixel shape (over its bounding box). The complete set of descriptors for each superpixel counts 2817 dimensions.

Ring interactions. We tested the proposed ring-based strategy by employing different radii from which learn urban-class interaction potentials. We evaluated the proposed strategy by considering 3 different radii for the SRP: 20, 40 and 50 meters. We will refer to these models as LEARNED SRP 20M, LEARNED SRP 40M and LEARNED SRP 50M. Regarding MRP, the set of radii of the concentric rings considered is $\rho_e = \{20, 40, 50\}$ meters and it is denoted as LEARNED MRP. Recall that the final inference over a test image will be performed over pairwise cliques composed only by adjacent superpixels (i.e. sharing some boundary).

[1] available at http://www.di.ens.fr/~aosokin/

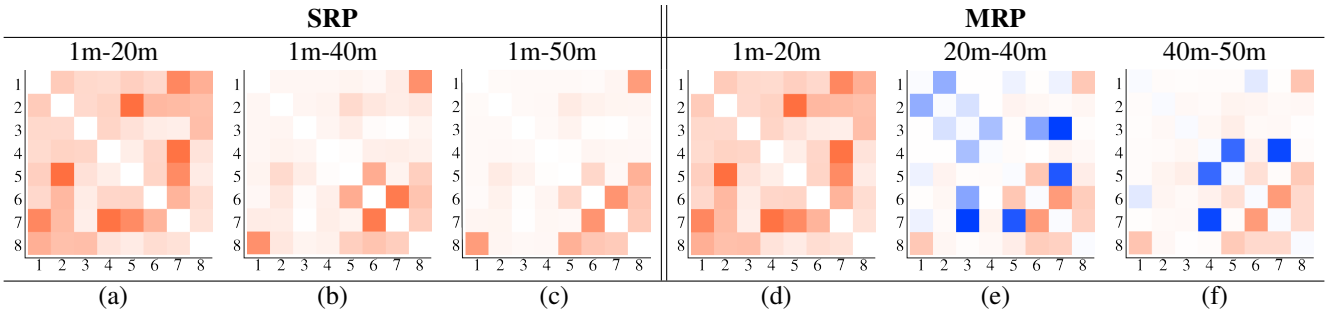


Figure 3. **Average class preference (pairwise weights).** Color ranges from blue (low weight), to red (large weight). White is the zero point. SRP (a)-(c) represent interactions learned inside the whole region, while MRP weights represent learned interaction for each ring.

Competing methods. We tested the max-margin discriminative ring-based learning of the CRF model versus different competitors of increasing expressive power. The first baseline is a nonlinear appearance classifier, built with random forests (RF UNARY) [5]. Random forest combined with the set of powerful descriptors is representative of the state-of-the-art in remote sensing literature [25]. This classifier takes as input the same 2817-dimensional set of descriptors employed by all other methods. Depth of the RF classifier is tuned by cross-validation over training images. We employed and ensemble of 200 trees, as cross-validation accuracy saturated for a larger number. As an additional baseline exploiting only appearance descriptors, we implemented a direct multiclass SVM [7], corresponding to a SSVM learning only the linear combination of descriptors for each class, without any pairwise interaction between nodes (MCSVM UNARY). The third approach implements a standard contrast-sensitive CRF (CS POTTS). We tested different similarity functions for the pairwise term, i.e. Gaussian kernel, spectral angle mapper and Chi-squared similarity, and we found experimentally that the latter performs better. It is computed as $g(x_i, x_j) = 1 - \chi(x_i, x_j)$, where $\chi(x_i, x_j)$ computes the Chi-squared distance between nodes i and j. Features x_i and x_j represent the ℓ_1-normalized spectral histograms, encoded in 21 bins per channels. The unary potential of this CRF is given by the negative log-likelihood for each superpixel based on the posterior probability given by the appearance classifier (RF). The trade-off parameter between unary and the contrast-sensitive pairwise potential is tuned by cross-validation on training images.

In order to evaluate the proposed approach versus another competitor able to account for local label co-occurrence from partial annotations, we augmented the contrast sensitive CRF mentioned above with a co-occurrence score between labels. To this end, for each superpixel we count the occurrences of labels of superpixels whose center is within a radius selected by cross-validation. Consequently, we can estimate the interaction potential as the label co-occurrence. We denote this approach Passive Ring Potentials (PASSIVE RP) and it is combined with the CS POTTS model. It can be seen as a single ring potential model with pairwise weights estimated as normalized counts from training data, instead of being learned by the SSVM. Again, the combination of the potentials is tuned by cross-validation over training images. The radius size providing best accuracy is 200m, cross validated in 10m steps between 20m and 300m. Note that cross-validation accuracy of the PASSIVE RP reaches a plateau from 60m on.

Evaluation. To evaluate test scores we performed a leave-one-out estimation. That is, we held one image out and trained the model on the remaining 19 scenes. We selected hyperparameters of the system (SSVM C and convergence rate, trade off-parameters of the competitors) on the 19 images by 3-fold cross-validation. Then, we infer a labeling on the held-out image.

We evaluate the proposed system on the accumulated held-out error matrix by employing 4 different criteria: mean average accuracy (AA), the estimated Cohen's Kappa statistic (κ) [6], the F1 score (F1) and per-class average accuracy. The first metric is the average of the per-class average accuracies. Each per-class accuracy is computed as the percentage of correctly segmented pixels over the total number of ground truth pixels, for this class. The second score (κ) is an overall accuracy metric which compensates for the chance agreement between classes. The third score is the average of the harmonic means between precision and recall for each class. This measure is sensitive to class-accuracy, but additionally takes into account the number of correctly classified pixels over the number of *predicted* labels for each class.

5.2. Results

Learned weights (Fig. 3). The models based on ring interactions count $D|\mathcal{Y}| + R|\mathcal{Y}|^2$ parameters, where D is the dimensionality of the descriptor set. The weight vector has size of 22600 for SRP and 22728 for MRP. Fig. 3 shows weights corresponding to discriminatively trained pairwise potentials. A low weight (towards blue) corresponds to a preference to particular class associations and consequently encouraged by the model. Larger weights (towards red) rep-

Table 1. Numerical results for the urban segmentation task. **Bold** and *italic* numbers refer to largest and second largest scores, respectively.

Model	AA	κ	F1	Roads	Build.	Trees	Grass	Soil	Water	Rails	Pools
RF UNARY	72.19	80.15	75.10	81.80	**87.04**	94.14	84.38	64.25	91.08	2.11	72.72
MCSVM UNARY	74.55	79.75	*76.43*	80.77	84.88	92.68	84.79	69.99	93.54	9.73	80.03
CS POTTS	73.71	79.24	76.04	*83.86*	86.41	*95.07*	82.88	67.10	91.65	3.46	79.22
PASSIVE RP	73.72	80.63	76.08	83.47	*86.68*	**95.10**	82.49	67.58	91.62	3.32	79.50
LEARNED RP 20M	76.82	**82.08**	**77.85**	83.83	86.39	94.04	86.58	71.32	93.55	16.99	81.87
LEARNED RP 40M	**78.35**	81.28	74.73	**84.02**	83.79	93.05	**86.92**	*74.53*	93.33	**21.35**	*89.77*
LEARNED RP 50M	76.62	80.65	71.76	83.79	83.07	92.10	86.65	73.93	*94.10*	*17.94*	81.39
LEARNED MRP	*78.07*	*81.61*	72.43	80.50	86.63	93.99	*86.72*	**75.51**	**94.31**	14.80	**92.13**

resent rare class combinations and therefore discouraged. White color corresponds to zero, thus a "neutral" class-association potential. These matrices are the average of the weights learned for each leave-one-out test. LEARNED SRP weights are shown in Fig. 3(a)-(c), and they show similar co-occurrence structures, shifting toward homogeneous weights as the radius grows. Differently to the strategy employed in PASSIVE RP, SSVM training finds no additional discriminant information in larger rings.

The LEARNED MRP weights show more interesting patterns. The first ring considering connectivity from 1m to 20m are very similar to the one learned by LEARNED 20M. For 20m-40m and 40m-50m rings, blue cells correspond to class association that are preferred over those involving a same class (on the diagonal). Although it may seem surprising, it is common that two pairs of classes are not occurring frequently when taken at the considered spatial buffer. This is particularly true for urban areas characterized by recurrent spatial pattern of classes. In the LEARNED SRP situation, same-class occurrences were always more frequent, since *all* labeled nodes closer than the ring size are considered. In the LEARNED MRP, for instance, at 20m - 40m level urban classes "Buildings" and "Roads" are co-occurring more frequently than "Buildings"-"Buildings" or "Roads"-"Roads". In this particular cases, the LEARNED MRP enforce *relative repulsion* to same-class situations, thus biasing spatial smoothing to particular urban class combinations. This means that LEARNED MRP go beyond simple label smoothness.

Quantitative results (Tab. 1). The basic CRF enforcing spatial smoothness (CS POTTS and PASSIVE RP) are providing equal or superior accuracy with respect to RF UNARY. The MCSVM UNARY outperformed the RF UNARY, suggesting that a class-specific learning of descriptors weights is an effective approach. This unary classifier performs overall very similarly to the CS POTTS.

Discriminatively trained CRF with ring-based potentials significantly outperform the baseline CRF tested, thanks to the joint learning of unary and pairwise potentials. The most accurate schemes are LEARNED SRP 20M and LEARNED SRP 40M, depending on the accuracy score employed. LEARNED SRP 20M provides better scores on the κ and F1, suggesting balanced omission and commission errors. However, by looking only at the average accuracy, LEARNED SRP 40M results the most accurate model. Regarding per-class scores, LEARNED SRP 40M wins in 3 out of the 8 classes and it is the second best model in 2 more cases. Interestingly, LEARNED SRP 20M wins only in terms of global accuracy metrics, but among ring potential-based method, provides higher accuracy on very locally structured urban classes such as "Buildings" and "Trees". The LEARNED SRP 40M provides in general good accuracy for larger and spatially homogeneous classes. We observe also that LEARNED SRP 50M does not perform better than other SRP models, suggesting that spatial information at 50m radius may not be discriminative enough.

Finally, the LEARNED MRP model offers a trade-off between the three LEARNED SRP models, showing the second best AA and κ scores. This is also reflected in the average accuracy for each class, offering best segmentation accuracy on 3 classes and second best on 1 out of the 8 urban classes considered, respectively. Consequently, we may obtain accuracies that are as good as the ones of SRP without manually specifying the optimal radius size. Details of the obtained urban segmentation maps are provided in Fig. 4.

6. Discussion and conclusions

In this paper, we proposed a system to train discriminatively CRF models for semantic segmentation of urban classes from high resolution satellite / aerial imagery. Specifically, we proposed two strategies for learning pairwise potentials based on ring structures to account for local class preferences. These potentials show two major improvements over standard contextual approaches: they can be trained from sparse but homogeneous annotations (a common situation in remote sensing image labeling problems) and, since they are trained discriminatively, they directly maximize the segmentation accuracy. The learned pairwise potentials consequently reflect the local co-occurrence of urban classes. The proposed joint learning of appearance and spatial interaction weights by the SSVM results in the best models.

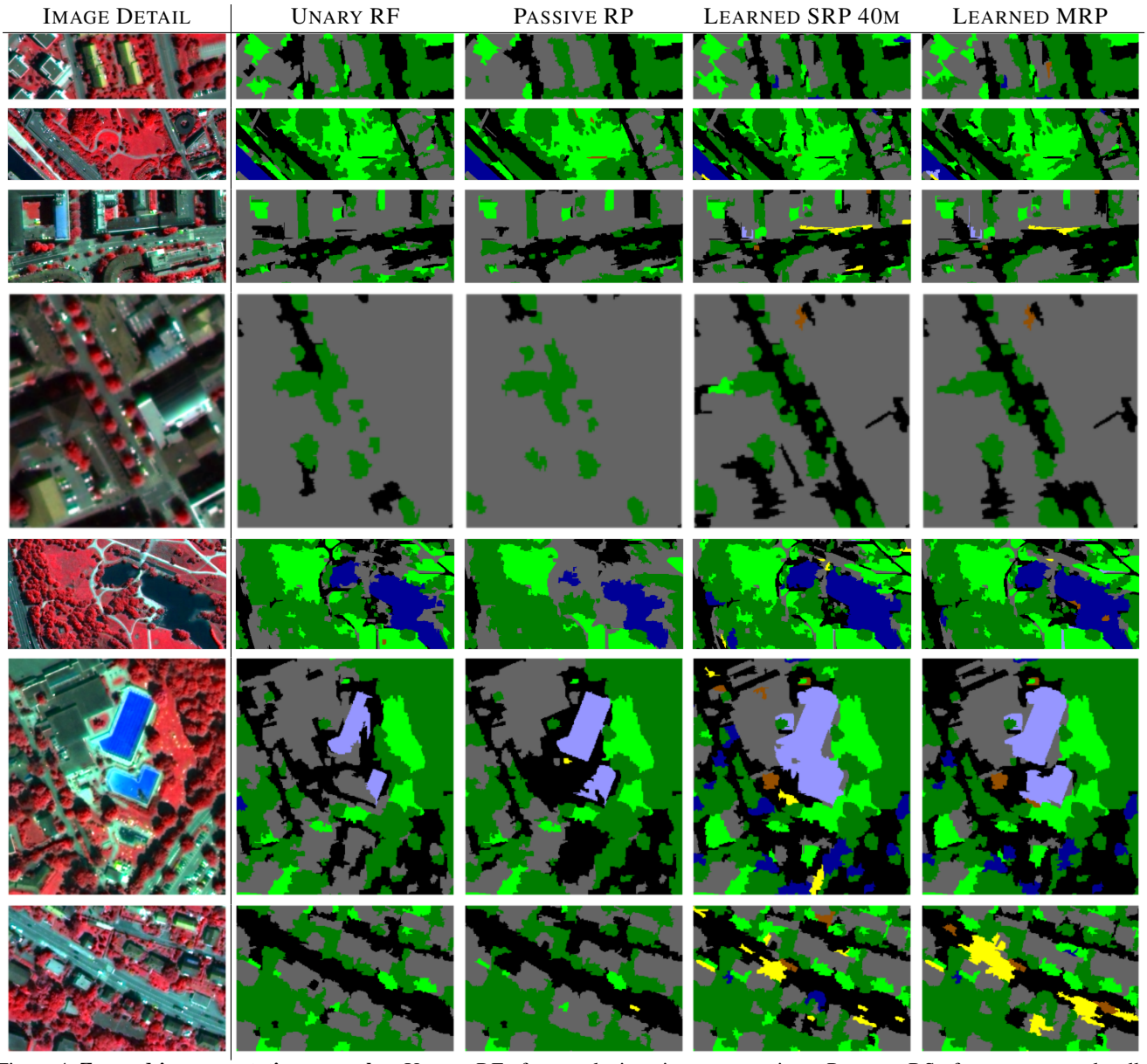


Figure 4. **Zoomed-in segmentation examples.** UNARY RF often results in noisy segmentatinos. PASSIVE RS often compesates locally by enforcing correct co-occurrences, but cannot cope with large errors of the unary. LEARNED SRP 40M and LEARNED MRP are able to successfully alleviate oversmoothing, at the cost of a less strict co-occurrence representation (see last row).

Urban segmentation accuracy is greatly improved with respect to a series of standard baselines with growing expressive power. The LEARNED SRP 40M model provides the highest accuracy. However, it hard-codes the spatial dependency by a circle with fixed radius. The second model proposed (LEARNED MRP) learns class preferences on a spatial quantization offered by a series of concentric rings. In this case, the importance of the each radius size in implicitly encoded in weights' values. Consequently, this model is more flexible and adaptive to the data at hand.

Acknowledgments

Michele Volpi was funded by the Swiss National Science Foundation Grant No. P2LAP2-148432 (http://p3.snf.ch/project-148432). Vittorio Ferrari was partially funded by the ERC Starting Grant VisCul.

References

[1] G. P. Asner, D. E. Knapp, E. N. Broadbent, P. J. C. Oliveira, M. Keller, and J. N. Silva. Selective logging in the brazilian amazon. *Science*, 310:480, 2005. 1

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36(2):192–236, 1974. 1

[3] E. Boros and P. L. Hammer. Pseudoboolean optimization. *Discrete Applied Mathematics*, 123(1-3):155 225, 2002. 4

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 3, 4, 5

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 6

[6] R. G. Congalton and G. Kass. *Assessing the Accuracy of Remotely Sensed Data*. CRC Press, 2008. 6

[7] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. 6

[8] P. Felzenswalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 5

[9] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE CVPR 2008, Anchorage (USA)*, 2008. 2

[10] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008. 1, 2

[11] M. Guillaumin, L. Van Gool, and V. Ferrari. Fast energy minimization using learned state filters. In *IEEE CVPR 2013, Portland (USA)*, 2013. 1, 2

[12] T. Hoberg, F. Rottensteiner, R. Queiroz-Feitosa, and C. Heipke. Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):659–673, 2015. 1, 2

[13] S. Kluckner, T. Mauthner, P. M. Roth, and H. Bischof. Semantic classification in aerial imagery by integrating appearance and height information. In *ACCV 2009, Xián (China)*, 2009. 1, 2

[14] V. Kolmogorov and C. Rother. Minimizing non-submodular functions with graph cuts – a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1274–1270, 2007. 3, 4, 5

[15] L. Ladický, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV 2010, Hersonissos (Greece)*, 2010. 1, 2

[16] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML 2001, Williamstown (USA)*, 2001. 1

[17] A. Lucchi, Y. Li, X. Boix, K. Smith, and P. Fua. Are spatial and global constraints really necessary for segmentation? In *IEEE ICCV 2011, Barcelona (Spain)*, 2011. 1, 2

[18] G. Moser, S. B. Serpico, and J. A. Benediktsson. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101(3):631–651, 2013. 1, 2

[19] S. Nowozin, P. V. Gehler, and C. H. Lampert. On parameter learning in CRF-based approaches to object class image segmentation. In *ECCV 2010, Hersonisson (Greece)*, 2010. 1, 5

[20] S. Nowozin and C. H. Lampert. Structured learning and prediction in computer vision. *Foundation and trends in computer graphics and vision*, 6(3-4):185–365, 2011. 2

[21] A. Rabinovich, A. Vedaldi, G. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE ICCV 2007, Rio de Janeiro (Brasil)*, 2007. 2

[22] K. Schindler. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4534–4545, 2012. 2

[23] K. C. Seto, B. Güneralp, and L. R. Hutyra. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16083–16088, 2012. 1

[24] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. *International Journal of Computer Vision*, 102(2):329–349, 2013. 2, 3

[25] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler. Features, color spaces, and boosting: New insights on semantic classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):280–295, 2015. 1, 6

[26] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005. 3, 4

[27] D. Tuia, J. Muñoz Marí, M. Kanevski, and G. Camps-Valls. Structured output SVM for remote sensing image classification. *Journal of signal processing systems*, 65(3):301–310, 2011. 1, 2

[28] F. Tupin and M. Roux. Markov random field on region adjacency graph for the fusion of sar and optical data in radargrammetric applications. *IEEE Transactions on Geoscience and Remote Sensing*, 43(8):1920–1928, 2005. 2

[29] A. Vedaldi. A MATLAB wrapper of SVMstruct. http://www.vlfeat.org/ vedaldi/code/ svm-struct-matlab.html, 2008. 5

[30] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler. A higher-order CRF model for road network extraction. In *IEEE CVPR 2013, Portland (USA)*, 2013. 1, 2

[31] P. Zhong and R. Wang. A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. *IEEE Transaction on Geoscience and Remote Sensing*, 45(12):3978–3988, 2007. 2